Predicting September sea ice: Ensemble skill of the SEARCH Sea Ice Outlook 2008–2013

Julienne Stroeve
University of Colorado

Lawrence C. Hamilton
University of New Hampshire, Durham, lawrence.hamilton@unh.edu

Cecilia M. Bitz
University of Washington

Edward Blanchard-Wrigglesworth
University of Washington

Follow this and additional works at: https://scholars.unh.edu/soc_facpub

Part of the Sociology Commons

Recommended Citation

This Article is brought to you for free and open access by the Sociology at University of New Hampshire Scholars' Repository. It has been accepted for inclusion in Sociology Scholarship by an authorized administrator of University of New Hampshire Scholars' Repository. For more information, please contact nicole.hentz@unh.edu.
Predicting September sea ice: Ensemble skill of the SEARCH Sea Ice Outlook 2008–2013

Julienne Stroeve1,2, Lawrence C. Hamilton3, Cecilia M. Bitz4, and Edward Blanchard-Wrigglesworth4

1National Snow and Ice Data Center, Boulder, Colorado, USA, 2Also at Centre for Polar Observation and Modelling Pearson Building, University College London, London, UK, 3Department of Sociology, University of New Hampshire, Durham, New Hampshire, USA, 4Department of Atmospheric Sciences, University of Washington, Seattle, Washington, USA

Abstract Since 2008, the Study of Environmental Arctic Change Sea Ice Outlook has solicited predictions of September sea-ice extent from the Arctic research community. Individuals and teams employ a variety of modeling, statistical, and heuristic approaches to make these predictions. Viewed as monthly ensembles each with one or two dozen individual predictions, they display a bimodal pattern of success. In years when observed ice extent is near its trend, the median predictions tend to be accurate. In years when the observed extent is anomalous, the median and most individual predictions are less accurate. Statistical analysis suggests that year-to-year variability, rather than methods, dominate the variation in ensemble prediction success. Furthermore, ensemble predictions do not improve as the season evolves. We consider the role of initial ice, atmosphere and ocean conditions, and summer storms and weather in contributing to the challenge of sea-ice prediction.

1. Introduction

The modern satellite record that began in the late 1970s has been invaluable in documenting climate change in polar regions. One of the most striking features is the large reduction and thinning in the Arctic’s floating sea-ice cover. Trends in the total sea-ice extent are negative during all calendar months [e.g., Serreze et al., 2007] but are largest at the end of the summer melt season in September. The mean September sea-ice extent declined by about 14% per decade, or 40% overall, from 1979 through 2013. Accompanying the extent reduction has been an overall thinning of the ice pack [e.g., Kwok et al., 2009], largely a result of first-year sea ice replacing the generally thicker multiyear ice [e.g., Maslanik et al., 2011]. With shrinking ice has come a rising demand for predictions at weekly to seasonal time scales, which have importance for ecosystems, coastal communities, marine access, and resource extraction.

All climate model simulations in the latest Coupled Model Intercomparison Project phase 5 indicate that the Arctic will eventually lose its summer ice cover as the concentrations of atmospheric greenhouse gases increase [e.g., Stroeve et al., 2012a], but studies also suggest that until this happens, the summer ice cover is likely to become more variable from year to year. For example, Holland et al. [2010] demonstrate that the standard deviation in the September extent increases as the ice cover thins. This complicates the prospects for near-term ice forecasts.

Since 2008, a Sea Ice Outlook (SIO) organized by the Study of Environmental Arctic Change (SEARCH) has solicited predictions of the mean September sea-ice extent from the Arctic research community. This effort, together with the Sea Ice for Walrus Outlook, provides a forum for the international sea-ice prediction and observing community to compare their ideas.

Individual predictions are based on a variety of modeling, statistical, and heuristic approaches. They are solicited in three cycles each year, at lead times from 2 to 4 months—around the first of June, July, and August. The July SIO predictions, for example, could be based on data through the end of June and are published in early July. The SEARCH website describes the background and rules for this enterprise [Study of Environmental Arctic Change (SEARCH), 2013].

In some years, the diverse methods used by the SIO contributors demonstrate collective skill in forecasting the September mean extent. In other years, they do not. Below we explore this history by comparing the distributions of SIO predictions with their target, the actual extent observed in September of each year. Our goal is not to evaluate individual predictions or particular methods. Rather, we analyze the SIO
predictions as an ensemble, which reveals a strong pattern. That pattern highlights the challenges for improved sea-ice prediction.

2. Overview of SIO and Contributions

The SEARCH SIO effort emerged from the discussions at the “Arctic Observation Integration Workshops” held in March 2008 in Palisades, NY, following the drastic and unexpected sea-ice decline witnessed in 2007. Each month during the summer melt season, a request is sent to the International Arctic Science Community soliciting predictions of the September ice extent. Submissions are reviewed by the Sea Ice Outlook Core Integration Group and Advisory Group and summarized in monthly reports that synthesize expectations from a broad range of prediction strategies [SEARCH, 2013]. The process starts in May and goes through September, culminating in a retrospective analysis after the season ends.

From 2008 to 2013, the SIO received 309 individual contributions—11 to 23 predictions each month (published in early June, July, and August) for 6 years. Many involve regression-type statistical models, estimated from historical data, then applied to forecast the near future. The predictors used in regressions might include sea-ice conditions (e.g., concentration, extent, and ice type), ocean temperatures, and atmospheric conditions (e.g., temperature, sea level pressure, and cloud forcing). For example, Drobot [2007] uses historical information on sea-ice concentration and fraction of multiyear ice, together with downwelling longwave radiation, surface temperature, and albedo as predictors. Lindsay et al. [2008] test the observed and modeled sea-ice and ocean conditions to improve forecast skill. Results suggest that for lead times of 2 months or less, sea-ice concentration is the most important predictor, but that for longer lead times, ocean temperature and sea-ice thickness become more important.

Alternatively, similar predictors could set initial conditions via data assimilation in forecasting techniques that rely on coupled sea-ice-ocean-atmospheric models and heuristic approaches. Some techniques use ensemble simulations from coupled ice-ocean models with prescribed atmospheric forcing from the historical record [Zhang et al., 2008a; Kauker et al., 2008], while others are coupled ocean-atmosphere-sea-ice models, initialized through data assimilation of ocean and sea-ice conditions. Ensemble members are typically constructed to sample the uncertainty in the forecast that is a result of intrinsic summer atmospheric variability.

Figure 1a graphs the monthly distributions of predictions for the September mean sea-ice extent from all the 309 SIO contributions over 2008–2013. The median SIO predictions are close to the observed ice extent in 2008, 2010, and 2011—years that roughly follow the longer-term trend. In 2009 and 2012, on the other hand, the median predictions are far off, and the observed ice extent falls well above or below any of the predictions. In 2013, the observed extent exceeds almost all of the predictions. These high-error years are the most anomalous in the observations since the SIO began, deviating both from the trend and the previous year. Figure 1b highlights the bimodal pattern of success or failure by graphing the median and approximate interquartile range (middle 50%) of the July predictions, together with the ensemble prediction error defined as distance from the SIO median to the observed September ice extent.

In later years, the SIO contributors were additionally asked to supply estimates of their uncertainty. They did so using varied methods from statistical confidence intervals to standard deviations, interquartile ranges, or other calculations, which are not formally comparable. Figure 2 draws informal comparisons by graphing the uncertainty bounds supplied by 15 contributors in July 2012 (Figure 2a) and 16 in July 2013 (Figure 2b), subdivided according to the SIO criteria into the broad categories of “modeling” or “statistical” methods. Other contributions in these months, including some classified as “heuristic,” did not provide uncertainty estimates. The observed 2012 ice extent, 3.63 million km$^2$, lies outside the intervals given with 11 of the 15 predictions and barely inside the lower limit of three others. As a group, the 2012 statistical predictions (median 4.35 million km$^2$) came closer to the unexpectedly low ice extent than the modeling predictions did (median 4.7 million km$^2$).

The observed ice extent was much lower than predicted in 2012, but in 2013, it was much higher (Figure 2b). The observed September 2013 ice extent, 5.35 million km$^2$, lies outside the intervals given with 13 of the 16 predictions. In 2013, the modeling predictions (median 4.35 million km$^2$), as a group, came closer than the statistical predictions (median 3.9 million km$^2$). The next section tests for significant differences among the SIO method types.

Figures 1 and 2 show a pattern of collective prediction success in years along the overall downward trend and collective failure when the observed extent was abruptly higher or lower. The same pattern identifying
Figure 1. (a) SEARCH Sea Ice Outlook predictions for June, July, and August reports compared with the observed mean September ice extent, 2008–2013. (b) Median and interquartile range of the July SIO predictions compared with the observed mean September extent.

Figure 2. Predictions with contributor-supplied uncertainty from Sea Ice Outlook in (a) July 2012 and (b) July 2013, shown with the observed September ice extent (gray line).
particular years as difficult to predict (2009, 2012, and 2013) occurs in two other less formal collections of sea-ice predictions from science agency office pools. Those two data sets are described and analyzed in the supporting information, using methods similar to Figure 1b.

3. Analysis of Prediction Errors

Figure 3a graphs the observed mean September extent for 1979–2013, along with the median July SIO predictions for 2008–2013. The trend in observations is summarized by a Gompertz curve—an asymmetrical S curve appropriate for the accelerating downward trend. If extended, this curve would approach zero asymptotically, instead of steepening toward zero as quadratic or exponential curves do (although all three fit well to the observations through 2013). This gives a reasonable approximation for the nonlinear historical trend, with residuals reflecting interannual variation. Figure 3b graphs the prediction errors (observed extent minus median July SIO prediction) against residuals from the Gompertz curve. SIO prediction errors and Gompertz curve residuals have a strong positive correlation ($r = 0.90, p < 0.05$). Ensemble prediction errors are largest in 2012 and 2013, the 2 years that depart most sharply from the trend.

Table 1 analyzes the SIO prediction errors in more detail, summarizing four quantile regressions in which the conditional median of absolute prediction errors is modeled as a function of year, month, and type of method used. Quantile regression provides a multivariate generalization of our median-based analysis in Figure 1, with similar advantages for these skewed and outlier-prone distributions (supporting information). Unlike ordinary least squares, quantile regression has high resistance to outliers and does not assume normality [Hamilton, 2013]. We also cannot assume that disturbances are independent and identically distributed, because the 309 SIO contributions represent about 91 different researchers or teams. Consequently, robust standard errors (Huber–White sandwich method) are employed for significance tests.

The coefficients in Table 1 are differences between the median absolute prediction errors for a given category of each variable (year, month, or method) compared with an arbitrarily-selected “base” category of that variable. For example, we chose 2011 as the base year, because the median prediction error is lowest in that year. Consequently, the coefficient for 2008 in Regression 1 (where all methods and months are combined) is

![Figure 3. (a) Observed September extent shown with Gompertz curve and median July SIO predictions and (b) prediction errors versus deviation from curve.](image-url)
0.34, indicating that the median absolute SIO prediction error for 2008 is 0.34 million km$^2$, greater than for the base year, 2011 (0.57 versus 0.23 million km$^2$). The standard error of this difference is 0.13 million km$^2$, so the median error in 2008 is significantly larger than in 2011 ($p < 0.05$). In Regressions 2, 3, and 4, each year coefficient reflects adjustments for other variables (i.e., month or method). Regression equations are written out in the supporting information.

Regression 1 is simply the regression of absolute prediction errors on (0,1) indicators for year. Median errors are largest for 2013, followed by 2012 and 2009. Regression 2 tests whether, adjusting for year, ensemble predictions improved from June (the base month) to July or August. Although predictions by some individual contributors improved, the ensemble performance did not: The median July errors are slightly larger than June (+0.06 million km$^2$) and in August, slightly smaller (−0.04 million km$^2$). Neither difference is significant.

Regressions 3 and 4 compare method types, with “heuristic” as the base method. A distinction among methods was not made in 2008, the first SIO year. For 2009–2013, both modeling and statistical approaches obtain lower median absolute errors than heuristic methods. Regression 3 employs the full 2008–2013 SIO data set. Statistical methods have median prediction errors of 0.14 million km$^2$ less than heuristic, a difference that falls short of significance. Regression 4 excludes 22 SIO contributions classified as “general public” (mostly from 2012 and 2013, when interest in the SIO process broadened). Among the remaining contributions, those based on statistical methods perform better than heuristic (−0.20 million km$^2$). Although modeling methods fare less well than statistical methods overall, they have a significant advantage in the unexpectedly high-extent year of 2013 (supporting information). These analyses confirm however that year-to-year conditions remain the dominant source of variation in ensemble prediction success.

### 4. Importance of Preconditioning

After the 2007 minimum, there was a growing consensus that the low extent in 2007 was largely a result of atmospheric forcing [e.g., L’Heureux et al., 2008; Kay et al., 2008; Schweiger et al., 2008; Zhang et al., 2008b]. In particular, a strong Arctic atmospheric dipole anomaly—that featured anomalously high sea level pressure

---

**Table 1.** Quantile Regression Coefficients With Robust Standard Errors, Modeling the Median Absolute Prediction Error as a Function of (0,1) Indicators for Year, Month, and Method$^a$

<table>
<thead>
<tr>
<th></th>
<th>Regression 1</th>
<th>Regression 2</th>
<th>Regression 3</th>
<th>Regression 4 (nonpublic)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Year</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td>0.34(0.13)$^b$</td>
<td>0.30(0.11)$^c$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>0.54(0.07)$^b$</td>
<td>0.46(0.09)$^d$</td>
<td>0.41(0.08)$^d$</td>
<td>0.42(0.09)$^d$</td>
</tr>
<tr>
<td>2010</td>
<td>0.24(0.10)$^b$</td>
<td>0.20(0.11)</td>
<td>0.10(0.10)</td>
<td>0.08(0.10)</td>
</tr>
<tr>
<td>2011</td>
<td>base</td>
<td>base</td>
<td>base</td>
<td>base</td>
</tr>
<tr>
<td>2012</td>
<td>0.54(0.09)$^d$</td>
<td>0.50(0.09)$^d$</td>
<td>0.52(0.09)$^d$</td>
<td>0.52(0.09)$^d$</td>
</tr>
<tr>
<td>2013</td>
<td>1.02(0.13)$^b$</td>
<td>0.98(0.13)$^d$</td>
<td>0.90(0.14)$^d$</td>
<td>0.90(0.12)$^d$</td>
</tr>
<tr>
<td><strong>Month</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>June</td>
<td>base</td>
<td>base</td>
<td>base</td>
<td>base</td>
</tr>
<tr>
<td>July</td>
<td>0.06(0.08)</td>
<td>0.05(0.07)</td>
<td>0.00(0.08)</td>
<td></td>
</tr>
<tr>
<td>August</td>
<td>−0.04(0.07)</td>
<td>−0.12(0.08)</td>
<td>−0.12(0.08)</td>
<td></td>
</tr>
<tr>
<td><strong>Method</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heuristic</td>
<td>base</td>
<td>base</td>
<td>base</td>
<td>base</td>
</tr>
<tr>
<td>Modeling</td>
<td>−0.03(0.08)</td>
<td>−0.08(0.08)</td>
<td>−0.20(0.08)$^b$</td>
<td></td>
</tr>
<tr>
<td>Statistical</td>
<td>−0.14(0.08)</td>
<td>−0.20(0.08)$^b$</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Constant</strong></td>
<td>0.23(0.06)$^d$</td>
<td>0.27(0.08)$^c$</td>
<td>0.39(0.08)$^d$</td>
<td>0.45(0.09)$^d$</td>
</tr>
<tr>
<td><strong>Estimation sample</strong></td>
<td>309</td>
<td>309</td>
<td>265</td>
<td>243</td>
</tr>
</tbody>
</table>

$^a$Notation: 2008 represents a dummy variable coded 1 if year = 2008, 0 otherwise; July represents a dummy variable coded 1 if month = July, 0 otherwise; modeling represents a dummy variable coded 1 if method = modeling, 0 otherwise; and so forth. See supporting information for details.

$^b$The $p < 0.05$.

$^c$The $p < 0.01$.

$^d$The $p < 0.001$ t tests using robust standard errors.
over the Beaufort Sea coupled with anomalously low sea level pressure over Eurasia [e.g., Wang et al., 2009]—persisted throughout summer 2007. This weather pattern produced a meridional wind anomaly that helped to transport ice away from the shores of Alaska and Siberia toward the pole and into the North Atlantic, advected warm air from the south, and gave rise to clear skies under the high pressure.

Several studies have focused on preconditioning of the ice cover prior to the 2007 minimum [e.g., Stroeve et al., 2008; Lindsay et al., 2009]. This intuitively makes sense because under a thick ice regime, a summer circulation pattern favorable for ice loss may translate into large changes in ice volume, but not necessarily large changes in ice extent as the ice will still be thick enough to survive. It is possible to follow ice age through Lagrangian tracking of individual ice parcels [Fowler et al., 2004]. Older ice tends to be thicker [e.g., Maslanik et al., 2007], so changes in the overall age of the ice also imply changes in ice thickness. As more open water in September has led to more first-year ice in spring, preconditioning in the form of a larger fraction of thin first-year ice increases the likelihood of low summer sea-ice extent. The fraction of first-year ice in March correlates with the September sea-ice extent \( r = -0.75 \). That is largely explained however by their common downward trend (the “trends” described here are based on 35 years of data, 1979–2013, rather than just the six SIO years). The near absence of a March–September correlation in detrended series reemphasizes the importance of summer atmospheric and oceanic variability. We also see a correlation between the simulated mean March sea-ice volume, estimated by the Pan-Arctic Ice Ocean Modeling and Assimilation System [Zhang and Rothrock, 2003], and the September extent \( r = -0.86 \), but this too is largely explained by a shared nonlinear trend (supporting information).

The May ice extent was similar in 2012 and 2013, and there was a larger fraction of the first-year ice in 2013 at the start of the melt season. Despite these preconditions, the September extent was 1.74 million km\(^2\) higher in 2013 than in 2012. The cooler summer of 2013 (June–July–August air temperatures 1–3°C colder than in 2012 over the Arctic Ocean and 1–2°C colder than the 1981–2010 mean) reduced melting, which was not predicted by the SIO contributors. Modeling studies support a large role for the summer ice melt [Zhang et al., 2008a]. The extent to which the atmospheric conditions depend on the sea-ice conditions will further affect predictions. Only fully coupled atmosphere-ocean-sea-ice models run in an ensemble prediction mode can predict this atmospheric response to sea-ice conditions. In contrast, coupled ocean-sea-ice models with prescribed atmospheric conditions are not able to predict these interactions and therefore may suffer. However, there is no indication yet of fully coupled models yielding better sea-ice predictions among the SIO contributions.

The rapidly changing Arctic also complicates statistical predictions based on historically observed relationships. Using eight ensemble members from the National Center for Atmospheric Research (NCAR) Community Climate System Model version 3, Stroeve et al. [2012b] report that the detrended September ice extent is better correlated with the detrended March ice thickness as the ice cover thins. This result was also found by Holland and Stroeve [2011]. Yet the long-term predictive capability for the September minimum actually decreases [Holland et al., 2010]. The reduced predictive skill as the winter ice cover thins has been noted in some of the contributions to the SIO and appears to be coincident with the rapid thinning of the ice cover.

In summary, while a thin winter ice cover suggests that the September extent will not return to the levels seen in the 1980s and 1990s, it has not shown good predictive skill for year-to-year variation. At present, however, many prediction strategies are focused on state of the ice cover prior to the summer melt season and suggest that assimilating sea-ice thickness and concentration could improve seasonal forecasts.

5. Discussion

In some years, the SIO ensembles accurately predict the September mean extent, while in other years, the observed extent falls outside the range of any prediction. This is true regardless of the general method used for prediction and whether or not we exclude contributions classified as general public. The predictions tend to be poor when the sea ice departs from the long-term trend. Indeed, the root mean square error (RMSE) of SIO predictions is only slightly better than a series of linear trend predictions, each calculated from data up to but not including the target year. Both SIO and linear trends improve substantially on climatology; however, RMSE = 0.77 for SIO and 0.80 for linear versus 1.91 for climatology (also calculated from data up to but not including the target year).
Departures from trends reflect (1) synoptic conditions during the summer, (2) early spring sea-ice conditions, and (3) methods used to make the prediction. Since forecasts cannot accurately predict (1) beyond a week or two, we only focus on (2) and (3). Results shown here suggest that sea-ice thickness and/or age do not yet provide enough predictive skill to outweigh the importance of summer atmospheric conditions. Nor do the results establish that one general approach to sea-ice prediction consistently outperforms others.

The 6 year SIO period is very short for comparing predictions and observations. However, several groups have published hindcasts of sea-ice extent for at least a few decades, and they find that the correlations for detrended hindcasts at 4 month lead times are as high as about 0.6 [e.g., Chevallier et al., 2013] and in other studies are much lower [e.g., Lindsay et al., 2008; Sigmond et al., 2013]. Meanwhile, studies using a “perfect model” framework, in which ensemble integrations are initialized from a reference model integration, give evidence of initial value predictability for 1–2 years [e.g., Blanchard-Wrigglesworth et al., 2011; Tietsche et al., 2013]. Such studies neglect errors from imperfect knowledge of the initial state and therefore give the upper limit of predictability for a given model. Taken together, these results suggest that the SIO activity has the potential to provide skillful forecasts and that the quality of the initial conditions and the method in which they are utilized is a key area for improvement.

While the results shown by the SIO seem to indicate that extreme years are less predictable than nonextreme years, it is unclear whether this is a robust feature of the natural system or a result of noise (given the shortness of the time series—only 6 years). Tietsche et al. [2013] explicitly examined the predictability of extremely low September sea-ice events in a perfect model context and found a significant skill that beat both a climatological and damped persistence forecast from a 1 May initialization (equivalent to many of the June SIO submissions). Indeed, their September sea-ice extent forecast error of ~0.5 million km² compares favorably to the SIO errors of ~1 million km² in 2012 and ~1.3 million km² in 2013.

Nevertheless, how intrinsic predictability varies from one year to the next is poorly understood and subject of current research. One may draw lessons from synoptic meteorology, where it is well known that different synoptic situations have varying degrees of predictability, to hypothesize that some years could offer significantly higher predictability than others. Additionally, modeling studies show that different extreme events are the result of different forcings [Cullather and Tremblay, 2008] and thus likely have different levels of predictability.

We have not tried to analyze individual SIO contributions, as these compare with each other or by month and year. Such evaluation is left to the researchers, who can apply detailed understanding of their own methods and inputs. Certainly some approaches rest on better grounded methods, and some show more skill in particular years. In contrast, others are presented with little or no science justification. The individual variations occur however within strong overarching patterns: for hard-to-predict years, the observed ice extent falls outside the range of any (or almost any) individual point predictions (Figure 1) and outside the uncertainty limits supplied for most (Figure 2). More nuanced classification of contributor approaches could refine this picture but leave the general findings to date unchanged.

Acknowledgments
This research was carried out under the Sea Ice Prediction Network project, with support from the U.S. National Science Foundation (PLR-1303938) and the Office of Naval Research (N00014-13-1-0793). Helen Wiggins (ARCUS) provided Sea Ice Outlook data; Walt Meier (NSIDC) and Jennifer Kay (NCAR) supplied the office pool data analyzed in the supporting information. Matthew Cutler assisted with dataset preparation.

The Editor thanks two anonymous reviewers for their assistance in evaluating this paper.

References


